Combining multiple structure and sequence alignments to improve sequence detection and alignment: Application to the SH2 domains of Janus kinases

Bissan Al-Lazikani*†, Felix B. Sheinerman†, and Barry Honig‡

Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032

Communicated by John Kuriyan, University of California, Berkeley, CA, October 30, 2001 (received for review August 14, 2001)

In this paper, an approach is described that combines multiple structure alignments and multiple sequence alignments to generate sequence profiles for protein families. First, multiple sequence alignments are generated from sequences that are closely related to each sequence of known three-dimensional structure. These alignments then are merged through a multiple structure alignment of family members of known structure. The merged alignment is used to generate a Hidden Markov Model for the family in question. The Hidden Markov Model can be used to search for new family members or to improve alignments for distantly related family members that already have been identified. Application of a profile generated for SH2 domains indicates that the Janus family of nonreceptor protein tyrosine kinases contains SH2 domains. This conclusion is strongly supported by the results of secondary structure-prediction programs, threading calculations, and the analysis of comparative models generated for these domains. One of the Janus kinases, human TYK2, has an SH2 domain that contains a histidine instead of the conserved arginine at the key phosphotyrosine-binding position, βB5. Calculations of the pK_a values of the βB5 arginines in a number of SH2 domains and of the βB5 histidine in a homology model of TYK2 suggest that this histidine is likely to be neutral around pH 7, thus indicating that it may have lost the ability to bind phosphotyrosine. If this indeed is the case, TYK2 may contain a domain with an SH2 fold that has a modified binding specificity.

The development of profile-based sequence alignment methods has led to major improvements in the detection of remote sequence relationships (see, e.g., refs. 1–4). However, for extremely low levels of sequence similarity, profile methods often fail. Because there are many examples of strong structural similarity in the absence of a detectable sequence relationship, a number of techniques have been developed that exploit structural information in the detection of remote homologs. Threading methods (see, e.g., ref. 5), which are purely structure-based, were developed with this goal in mind, and multiple sequence profiles have been generated directly from multiple structure alignments (see, e.g., ref. 6). Increasingly, novel ways have been found to combine structural and sequence information. These include the incorporation of structural information directly into sequence profiles (7–9) and the integration of sequence information into threading algorithms (4, 10, 11). This paper describes an approach that uses multiple structure alignments to merge profiles that have been generated from high-quality multiple sequence alignments. A Hidden Markov Model (HMM) then is generated from the merged alignment. The approach is related to previous work that has combined multiple structure and sequence information (e.g., ref. 9); however, there are a number of essential differences. These are due, in part, to the fact that our primary goal is not the detection of remote homologs but rather the generation of an optimal alignment for a homolog that already may have been detected.

The utility of the approach is demonstrated through its application to the SH2 domain family (reviewed in ref. 12). SH2 domains are composed of about 100 amino acids and consist of two α-helices packed against a central, antiparallel β-sheet. SH2 domains are present in a large number of signal-transduction proteins. They mediate intramolecular regulation and intermolecular protein—protein association by binding to specific motifs containing a tyrosine residue. In almost all known cases, phosphorylation of the tyrosine located within a specific peptide sequence on the target protein is a prerequisite for SH2 binding. The SH2 domain family includes members with as little as $\approx$15% pairwise sequence identity, which is a level at which pure sequence-based methods often generate erroneous alignments. One such example is provided by the SH2 domain of the Cbl adapter protein, which was identified as an SH2 domain only after its structure was solved (13). Cbl-SH2 thus provides an ideal test of both sequence-detection and sequence-alignment methods.

Janus kinases (JAKs) are a family of nonreceptor protein tyrosine kinases involved in signaling cascades initiated by various cytokines, interferons, and growth factors (14). There are four human JAK proteins: JAK1–3 and TYK2. JAKs share seven main regions of homology, termed JAK-homology domains 1–7 (JH1–7), numbered from the C to the N terminus (15). JH1 is the C-terminal protein kinase domain, and JH2 is a kinase-like domain whose precise function remains unclear (16). JH3–7 play a role in receptor interactions. There has been considerable uncertainty as to whether JAKs contain SH2 domains. An early description of the sequence of murine JAK2 mentions an "intriguing, albeit tenuous similarity" of the portion of the JH4 domains in JAKs with sequences of some SH2 domains (15). More recent sequence-search methods predict the presence of SH2 domains in JAKs (17), and a multiple sequence alignment obtained from CLUSTAL W (18) was reported. These results notwithstanding, the presence of SH2 domains in JAKs is not accepted universally (see, e.g., refs. 19–21). Moreover, the Pfam database (22) includes only JAK2 in the SH2 family alignment, whereas JAK1, JAK3, and TYK2 are not classified as SH2-containing. The SMART database (23, 24) includes regions (in JH3–4) of JAK1–3 and TYK2 as SH2 domains, but the alignment does not include the BG loop that forms a part of the peptide-binding surface and the last β-strand (βG) that follows it. A recent study (25) reported that secondary-structure predictions for human JAK1, JAK2, and JAK3 are consistent with the presence of SH2 domains in these proteins and that fold recognition servers suggest the presence of an SH2 domain in human JAK2 and JAK3 proteins.

Abbreviations: JAKs, Janus kinases; PDB, Protein Data Bank; 3D, three-dimensional; HMM, Hidden Markov Model.

*Present address: Inpharmatica Ltd., 60 Charlotte Street, London W1T 2NU, United Kingdom.

†B.A.-L. and F.B.S. contributed equally to this work.

‡To whom reprint requests should be addressed. E-mail: bh6@columbia.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

 www.pnas.org/cgi/doi/10.1073/pnas.011577898

Despite the rather strong evidence that JAKs contain SH2 domains, a number of questions remain. In addition to the absence of an alignment in the C-terminal region, the alignment for TYK2 suggests that if it contains an SH2 domain, it is a highly unusual one. In all reported multiple alignments (see, e.g., refs. 18, 24, and 25), there is a histidine present at βB5, a position that contains a conserved arginine in every known SH2 domain as well as predicted SH2 domains in JAKs 1–3. This arginine is in direct contact with the phosphotyrosine and makes an important contribution to the binding affinity of phosphotyrosine-containing peptides (12, 26). Indeed, substitution of this arginine to a lysine in the SH2 domain of Abl kinase completely abolishes the binding function of this domain (27). It is possible, of course, that substitution of the conserved arginine with another basic residue in other SH2 domains has less dramatic effects and that a charged histidine in TYK2 plays the same role as the arginine. This, in fact, has been implicitly assumed in reported alignments. Another possibility is that the putative SH2 domain of TYK2 does not bind peptides that contain phosphotyrosine or can do so only under specific conditions. Specifically, given that the pK_a values of histidines in proteins can vary over many pH units, it is not at all clear whether the His in TYK2 is charged and whether SH2-TUK2 has the capacity to bind phosphotyrosine.

We have used the sequence alignment obtained from our profile method as a basis for addressing these issues. The profile clearly predicts the existence of an SH2 domain in all JAKs, and the resulting alignment includes the entire structural domain. The alignment is used to generate homology models for all JAK-SH2s, and a number of tests suggest that these are viable models. In agreement with the conclusions of previous studies, a His is located at the βB5 position in TYK2. However, in contrast to previous assumptions, evidence is presented that this His is unlikely to be protonated at pH 7. Thus, TYK2 appears to contain a highly unusual SH2 domain. Possible implications of this finding are discussed.

Methods

Selection of Representative SH2 Structures. The Protein Data Bank (PDB) (28) contains more than 100 structures of SH2 domains. Of these, we selected 19 representative structures so that no two domains in the set have more than 95% sequence identity to one another. For structures with similar sequences (>95% sequence identity), those solved to highest resolution were selected. Where possible, structures determined with x-ray crystallography were chosen over those determined with NMR spectroscopy. The selected proteins and the associated PDB identifiers are listed in Fig. 1.

Structure-Based Multiple Alignment Procedure. A structure-based sequence alignment of the 19 nonredundant SH2 domains is produced by using the multiple structure superposition routine implemented in PRISM (6, 29), followed by some manual adjustments to the loop regions (see Fig. 1). A Position-Specific Scoring Matrix was generated with PSI-BLAST (2) based on the structure-based alignment and used to perform PSI-BLAST searches of the SwissProt database (30) for other SH2 domains. The searches were performed with the BLOSUM62 substitution matrix (31) by using gap initiation and extension penalties of 11 and 1, respectively. Nineteen searches, using each of the SH2 domains shown in Fig. 1 as a probe sequence, each biased by the Position-Specific Scoring Matrix, were carried out. Two hundred and two sequences matched with expectation values of 0.01 or better (excluding the JAKs) are all annotated as SH2 domains in SwissProt. These sequences were used to generate a sequence profile of the SH2 domain family.

Each of the 202 sequences was aligned to each of the sequences of the 19 SH2 domains of known structure. A multiple sequence alignment of all sequences that were greater than 50% identical to a given PDB sequence (over a region of at least 70 residues in length) was carried out with CLUSTAL W (18). This resulted in 19 separate multiple sequence alignments. The 19 multiple sequence alignments then were combined based on the structure-based alignment shown in Fig. 1, yielding a single multiple sequence alignment containing a total of 141 sequences (including the 19 sequences shown in Fig. 1). This alignment is used to generate an HMM, which was used to align the SH2 domains that were not included in the earlier step. The HMM was built by using the HMMER 2.1.1 package (S. Eddy in http://hmmer.wustl.edu), and the alignment was carried out with the HMMALIGN utility in HMMER 2.1.1. In this process, the alignment of SH2 domains in the combined (containing 141 sequences) multiple sequence alignment was kept fixed. A flowchart of the procedure is given in Fig. 2. The sequence profile of the SH2 family generated as described then is used to construct an HMM and to align sequences of putative SH2 domains in JAKs.

The rationale for the procedure described in this section stems from an observation that, for closely related sequences (e.g., >50% identical), global sequence alignment performed with the Needleman—Wunsch method (32) implemented in the CLUSTAL W package produces high-quality alignments similar to those obtained from structural superposition, whereas the explicit use of structural information becomes crucial for the alignment of proteins with low sequence similarity (33, 34). This is illustrated in Fig. 3*a*, where the structure-based alignment of a subset of SH2 domains is compared with the sequence-only-based alignment of these proteins, reported in the Pfam database (22). As can be seen, for two of the four proteins shown, a significant fraction of the domain, including the C terminus α-helix, is missing in the Pfam alignment.

It is of interest to compare the alignment obtained from the flow chart in Fig. 2 with a more conventional, HMM-based alignment. To this end, we built an HMM based on the alignment of the 19 SH2 domains of known structure shown in Fig. 1 and used it to align all 202 sequences of SH2 domains found in SwissProt. The alignment produced in this way contained significantly more gaps, many of which fell within secondary structure elements of SH2 domains, e.g., 27 gaps are introduced within secondary structure of v-src SH2 (shown in red and yellow in Fig. 1), compared with the total of 7 gaps, introduced by the alignment procedure described in Fig. 2. Visual inspection also revealed that similar sequences were occasionally misaligned in the alignment of 202 SH2 sequences based on the HMM built on 19 superimposed structures.

Construction and Evaluation of Homology Models. The multiple alignment shown in Fig. 1 was used to build a homology model of TYK2-SH2 as well as of putative SH2 domains in other human JAKs and of other SH2 domains. The program MODELLER 4.0 (35) was used, and all 19 structures were used simultaneously as templates. The Verify3D server (36, 37) was used to assess the quality of each model. The server analyzes a protein in terms of the suitability of each residue to be found in its specific local environment. The properties evaluated include the buried surface area of the residue, the fraction of the side chain surface area in contact with polar atoms, and the local secondary structure. A database of the precomputed preferences for each residue type to be found in specific environments is used to calculate a score for each residue in the model, with a higher score corresponding to a more favorable environment (37).

Results

Test of the SH2 Profile-Based Alignment: Alignment of Cbl-SH2. As a test of the ability of the sequence profile generated for the SH2 family to align remote homologues, we aligned the sequence of the SH2 domain of the Cbl adapter protein (13) onto the profile. The sequence identity between Cbl-SH2 and other SH2 domains used in the structure-based alignment shown in Fig. 1 ranges from 7% to 20%. The presence of an SH2 domain in Cbl was not predicted

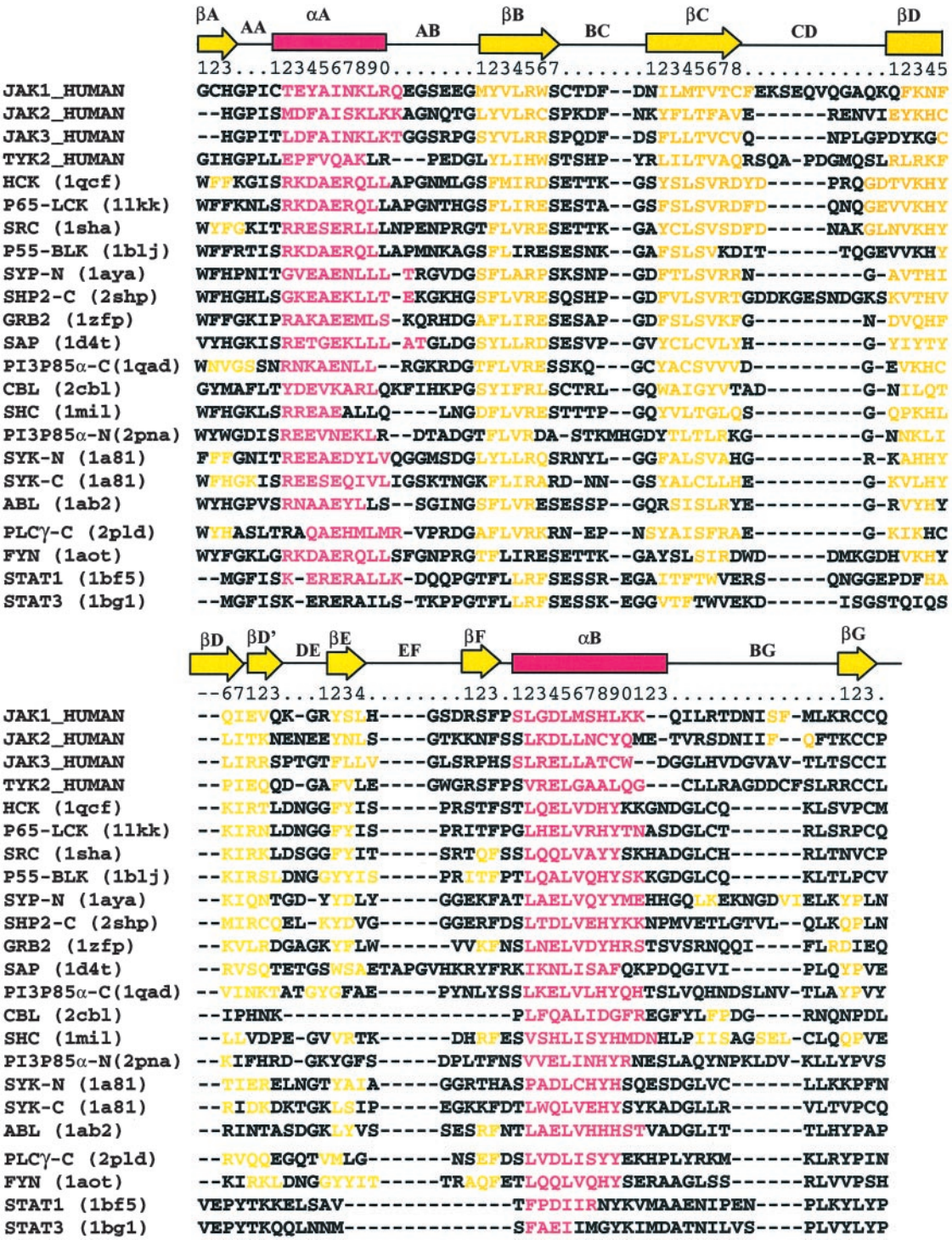

Fig. 1. Alignment of JAKs and 19 SH2 domains of known structures. The multiple alignment of the SH2 domains is structure-based (6), whereas the four JAKs are aligned with an HMM of the SH2 domain family (see text for details). Minor manual modifications were made to close gaps in predicted secondary structures, e.g., the αB helix of TYK2. The conventional secondary-structure assignments and numbering for SH2 domains (12) are illustrated in cartoon form above the alignment. Residues are colored according to secondary structure: gold, β-strand; magenta, α-helix. For the JAKs, the secondary structure displayed is that predicted by JPRED2 (39). For the 19 SH2 domains, the DSSP (49) secondary-structure assignments from the PDB coordinates are shown.

based on sequence comparisons (13, 38), and, in contrast to putative SH2 domains in JAKs, Cbl-SH2 is not detected with an expectation value below 0.01 in our PSI-BLAST searches (the best E value is 2.4). The structure of Cbl-SH2 was excluded from the structure-based profile used in this test. Given the low level of sequence identity, an accurate alignment of Cbl-SH2 with other SH2 domains represents a good test of the sensitivity of sequence-alignment procedures.

Fig. 3*b* reports alignments of Cbl-SH2 with Src-SH2 obtained from a number of methods. The structure-based alignment generated with PRISM is used as a standard. An alignment of Cbl-SH2

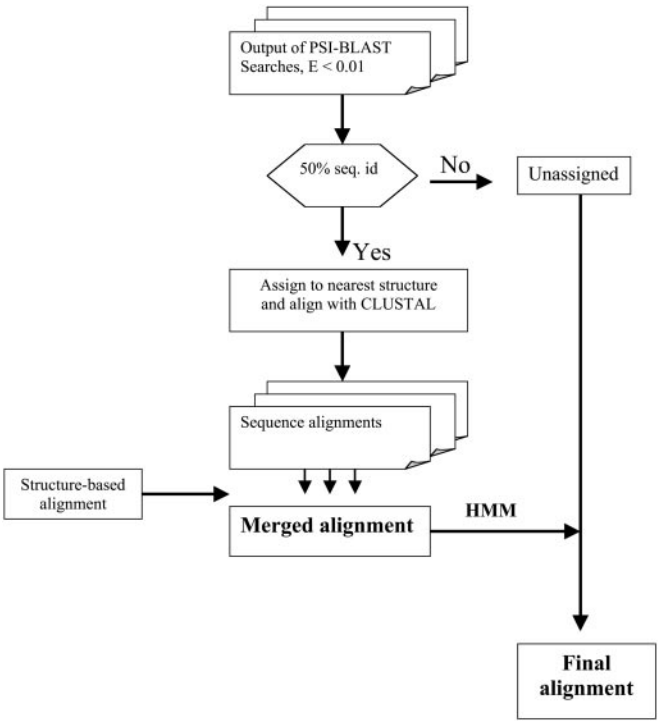

Fig. 2. Flowchart of the procedure used to align all SH2 domains.

with Src-SH2 based on an HMM built on a sequence profile of the SH2 family, generated as described in *Methods*, is quite similar to the one obtained from structural superposition. In contrast, the Needleman–Wunsch method that attempts to align the entire sequence fails completely whereas PSI-BLAST produces a reasonable alignment but for only part of the sequence.

Sequence-Based Identification of SH2 Domains in JAKs. The PSI-BLAST searches carried out as described in *Methods* identified TYK2 and JAK1–3 as SH2 domains with highly significant scores; the best E values obtained upon convergence range from 10^{-11} (for TYK2) to 10^{-20} (for JAK1 and JAK2). The JAK-SH2 sequences also were used as probe sequences to search the nonredundant database of the National Center for Biotechnology Information, nrDB (release of July 2000) by using PSI-BLAST (2) with default parameters. In all four searches, a large number of SH2 domains were matched with E values ranging between 2×10^{-12} and 2×10^{-27}.

Alignment of the full JAK1–3 and TYK2 sequences onto the SH2 sequence profile, performed as described in *Methods*, aligned the regions spanning residues 426–533, 399–500, 375–476, and 450–552 from human JAK1, JAK2, JAK3, and TYK2, respectively. A multiple alignment of the JAK SH2 domains with 19 sequences for which structures are available is shown in Fig. 1. It is important to note that the entire domain, including the βG strand and the preceding loop, is included in the alignment. As noted above, these regions were absent in the previously reported alignments.

Structural Analysis of Alignment Quality. The secondary structure for each of the JAKs was predicted by using the JPRED2 server (39). JPRED2 produces a consensus secondary structure as predicted by many programs such as PHD (40) and DSC (41). Fig. 1 shows the alignment of all four JAKs, with the 19 SH2 domains colored by secondary structure. The predicted secondary structures are in agreement with the secondary structures of typical SH2 domains. As can be seen in Fig. 1, the hydrophobicity patterns of each of the JAK-SH2s are also in very good agreement with the SH2 structures.

Fig. 3. Sequence- and structure-based alignments of SH2 domains. (*a*) Alignments of several SH2 domains of known structure (colors indicate secondary structure elements, as in Fig. 1). (*b*) Alignments of Cbl with Src-SH2. Residues within Cbl-SH2 are shown in bold.

Moreover, GENTHREADER (4) matched all putative JAK-SH2 domains with the structures of known SH2 domains with a confidence level of "certain."

A more detailed structural analysis of the sequence alignment reveals other features of SH2 domains that are present in JAKs. Residues βB2–4 are buried deeply in the SH2 structures listed in Fig. 1 and, therefore, are likely to be important for SH2 stability. These are aromatic–aliphatic–aliphatic residues in all 19 structures and are highly conserved among all SH2 sequences. They are aligned with Tyr-Val-Leu in JAKs 1–3 and Tyr-Leu-Ile in TYK2. The highly conserved Gly-AB7 adopts an unusual $+\phi$, $-\psi$ conformation and thus is likely to be important for the proper folding of an SH2 domain. The mutation of this residue to glutamate in Bruton's tyrosine kinase (Btk), found in patients with X-linked Agammaglobulinemia (XLA), introduces severe structural alterations in the SH2 domain. All four JAKs contain a glycine at this position. Three other specific mutations in patients also cause considerable perturbation of the structure of the Btk-SH2 domain

(42). These mutations are Tyr-βD5 to Ser, Leu-αB5 to Phe, and His-αB9 to Gln. Human JAK1–3 and TYK2 contain either Phe or Cys at position βD5, both of which are seen in SH2 domains of known structures (see alignment of Fig. 1). Position αB5 is occupied by a Leu in JAK1–3 and TYK2, which is the amino acid found in most other known SH2 domains (Fig. 1). Position αB9 is occupied by aromatic and hydrophobic residues in JAK1–3 and TYK2 (Leu, Tyr, Cys, and Leu, correspondingly; see Fig. 1). Aromatic residues (Tyr, Phe) are also seen at this position in other SH2 domains.

The alignment we obtain for the BG loop is of particular interest because this region has not been included in previous JAK alignments. Examination of known SH2 domain structures reveals that, although the length of the BG loop is quite variable, there is a conserved BG loop "anchor" at the BG13 position. Most SH2 domains contain a leucine at this position that packs against the C terminus of the αB helix. A leucine is also present in our alignment of the JAK SH2 domains. Similarly, an aromatic residue is present at the βF3 position that, in SH2 domains of known structure, often packs against the N terminus of the αB helix. It seems clear from this analysis and from previous work that JAK family proteins have SH2 domains that fold into a structure that is quite similar to that of other SH2 domains.

Is the Binding Site Conserved? The most dramatic difference between TYK2-SH2 and other SH2 domains is the identity of the residue at the βB5 position. TYK2 contains a histidine whereas all other SH2 domains, including the other JAK proteins, contain an arginine at the βB5 position. Although it is possible that the histidine plays the same role as the conserved arginine in all other SH2 domains, it is not clear that the histidine is charged in TYK2 under normal conditions. In the following section, we report the construction of a homology model of the putative SH2 domain of TYK2. Our goal is both to determine whether such a domain is likely to be stable and to consider its binding properties in greater detail.

Homology Model of TYK2. A homology model of the SH2 domain in TYK2 was built as described in *Methods*. All 19 structures listed in Fig. 1 were used as templates. The Verify3D (36, 37) server was used to assess the quality of the model. The three-dimensional (3D) profiles generated for the model and for all 19 structures are shown in Fig. 4*a*. The profile of TYK2-SH2 falls within the range seen for the 19 experimental structures and does not score below 0.0 at any point, characteristic of a good model. The cumulative 3D scores, calculated by summing the 3D scores at each position in the profile (36, 37), are shown in Fig. 4*b*. The cumulative score of TYK2-SH2 model is 34, within the range obtained for the 19 SH2 structures (29.4–52.4, see Fig. 4*b*). Luthy *et al.* (36) presented a plot describing the relationship between the 3D score and the length of the protein in experimentally determined structures. The score obtained by the TYK2-SH2 model, which is 103 residues long, ranks among the scores obtained by medium-resolution x-ray structures and NMR structures.

Although the model clearly is not accurate in all its details, the fact that it scores in the range seen for SH2 domains of known structure suggests that it is a reasonable approximation to the actual structure. Homology models for the JAK1–3-SH2 regions were built in the same manner as for TYK2. As seen from the data presented in Fig. 4*b*, all human JAK sequences fit the architecture of an SH2 fold quite well, as judged by Verify3D scores.

Binding of Phosphopeptides: Structure of a Putative Binding Site. It appears quite likely that the SH2 domains of JAKs 1–3 bind phosphotyrosine-containing peptides, as do all SH2 domains of known structure. This is not the case for TYK2. The factors that determine the pK_a values of groups in proteins have been discussed extensively (43–46). Briefly, a basic group that is partially removed from solvent will tend to have its pK_a lowered

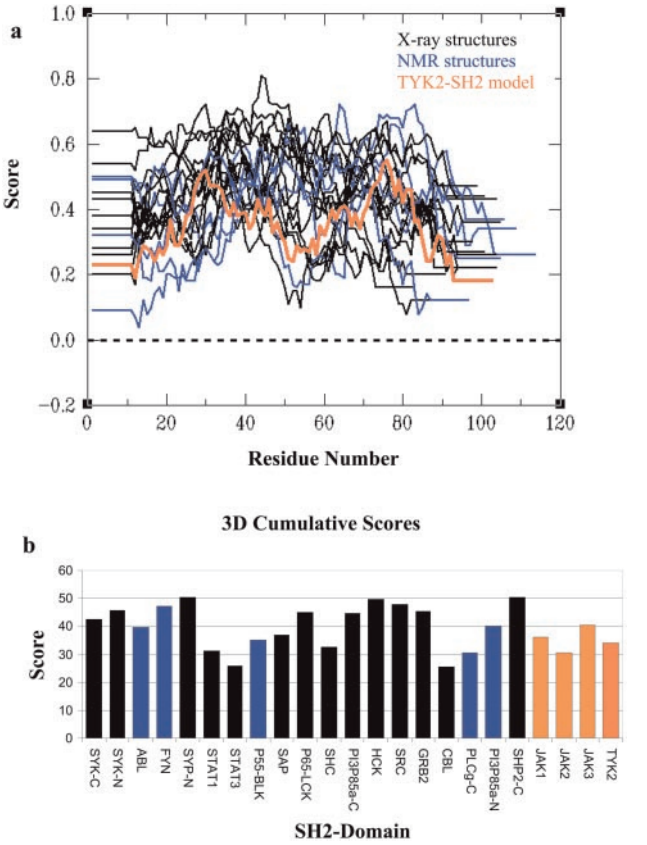

Fig. 4. 3D profiles and cumulative scores for JAK-SH2 models and 19 SH2 structures. Profiles were obtained from the Verify3D server (37). (*a*) 3D profiles of the 19 selected SH2 domain structures and of the T*yk*2-SH2 model. (*b*) Cumulative 3D scores for the 19 selected SH2 domain structures and of the four JAK models.

relative to the isolated amino acid because the charged form of the amino acid will be less well solvated in the protein than in water. The loss of aqueous solvation, in principle, can be compensated by stabilizing hydrogen-bonding interactions with polar groups in the protein or with negatively charged amino acids. Thus, the observed pK_a will depend on the local environment of the group in the protein. We have used the method of Alexov and Gunner (47) to calculate the pK_a values of arginines at the βB5 position in five different SH2 domains of known structure (v-src, SHPTP2-N, Cbl, PLCγ-C, and STAT1). All calculated pK_a values are quite high ($\geq$12) despite the fact that the partial burial of the Arg in the deep phosphotyrosine-binding pocket would, on its own, result in pK_a shifts to lower values. However, compensating interactions with nearby charged and polar groups raise the calculated pK_a values to those normally associated with Arg residues in proteins. The groups that make the largest stabilizing contributions are His or Gln residues at the βD4 position and/or spatially adjacent glutamic acids (at positions αA6 and BC1). His βD4 is, in particular, highly conserved perhaps because of its role in stabilizing the charge and position of the Arg.

The pK_a values of Arg βB5 in JAK1–3 are also predicted to be high. A histidine is present at the βD4 position in JAK2, and other groups interact favorably with Arg βB5 in JAKs 1 and 3. In contrast, His βB5 in our model of TYK2-SH2 is predicted to have an extremely low pK_a (near zero). This large shift results in part from desolvation effects and in part from the presence of a lysine at position βD4, where a stabilizing histidine is normally found. The only other SH2 domain with a highly basic residue at βD4 is the C-terminal SH2 domain in GAP, which contains an Arg at this

position. We built a homology model for this domain and found that the calculated pK_a of Arg-βB5 in the model of GAP-SH2 is above 13. This is due, in part, to strong stabilizing interactions with Asp-BC1. In contrast, TYK2-SH2 contains Thr at this position. To test the sensitivity of the calculated pK_a of His-βB5 in TYK2-SH2 to the structural model, pK_a calculations on the top 10 models generated by MODELLER were carried out. The average pK_a of His-βB5 in these models is 0.6, and the highest calculated pK_a is 2.3.

Although pK_a calculations on inaccurate models are unlikely to produce accurate values, the major point of the calculations reported in this section is the demonstration that His βB5 in TYK2 is unlikely to be protonated. As is the case for the arginines at this position in other SH2 domains, its presence at the bottom of a deep pocket inevitably will shift its pK_a to lower values. However, in contrast to other SH2 domains, the strongest interaction in this region (with-Lys βD4) only serves to further reduce its pK_a. The available evidence then suggests that TYK2-SH2 differs from all known SH2 domains in that, at around pH 7, it does not present a positive charge at the βB5 position.

Discussion

The alignment of protein sequences based on the 3D superposition of their structures is known to provide a means for protein-sequence comparison in cases in which similarities are weak and cannot be detected reliably by sequence-only methods. In this paper, we have used structural superposition to improve a multiple alignment in a case in which a weak sequence relationship already has been detected. The combined sequence/structure alignment approach should be applicable generally in cases in which a number of structures are available for members of a particular protein family. However, the appropriate strategy may well vary from family to family and depend on the number of available structures and their range of structural distances. A general-purpose algorithm that exploits the overall approach described in this work has been benchmarked extensively and yields significant improvement over existing Position-Specific Scoring Matrix methods (L. Xie and B.H., unpublished results).

As summarized above, a number of novel methods have been reported recently that combine sequence and structural information to improve the detection of remote homologs (4, 7–11). In contrast, the primary goal of this work has been to improve sequence alignments for family members that already have been identified, and this has dictated a somewhat different strategy. For example, we have used PSI-BLAST to detect these putative family members but have not relied on PSI-BLAST alignments. Rather, we have used pure sequence-based methods only for cases with high levels of sequence identity (50% in this paper), whereas more distant sequence neighbors are aligned only to the merged structure-based HMM. This procedure reduces the probability of errors in the multiple sequence alignment that result from alignment problems for distantly related sequences.

Our results provide strong support for previous studies (17, 25) that have concluded that JAK family proteins contain SH2 domains. The most surprising prediction of our study is that human TYK2 kinase contains an SH2 domain that cannot bind phosphotyrosine. The presence of a histidine instead of an arginine in the crucial βB5 position indicates that this is a unique SH2 domain. In principle, it is possible, of course, that the histidine simply replaces the arginine as a determinant of binding affinity. However, pK_a shifts from desolvation effects and the apparent absence of interactions that stabilize the charged form of the His argue that it is unlikely to attract a negatively charged substrate such as phosphotyrosine-containing peptides. A number of conjectures suggest themselves. It is possible that another residue in the binding site, specifically, Lys-βD4, coordinates a phosphotyrosine or that other still-unknown factors (for example, another protein domain or low pH) serve to enhance binding. Alternate possibilities are that TYK2-SH2 associates with a completely different class of targets and that its activity is not controlled by phosphorylation or, possibly, that it binds nonphosphorylated tyrosines. Resolution of these questions can come only from experimental studies.

This work was supported, in part, by National Institutes of Health Grant GM-30518 (to B.H.) and a Sloan/Department of Energy Postdoctoral Fellowship in Computational Molecular Biology (to F.B.S.).

1. Krogh, A., Brown, M., Mian, S. I., Sjolander, K. & Haussler, D. (1994) *J. Mol. Biol.* **235,** 1501–1531.
2. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
3. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol.* **6,** 361–365.
4. Jones, D. T. (1999) *J. Mol. Biol.* **1999,** 797–815.
5. Fischer, D. & Eisenberg, D. (1996) *Protein Sci.* **5,** 947–955.
6. Yang, A.-S. & Honig, B. (2000) *J. Mol. Biol.* **301,** 691–711.
7. Rice, D. W. & Eisenberg, D. (1997) *J. Mol. Biol.* **267,** 1026–1038.
8. Hargbo, J. & Elofsson, A. (1999) *Proteins* **36,** 68–76.
9. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000) *J. Mol. Biol.* **299,** 499–520.
10. Panchenko, A., Marchler-Bauer, A. & Bryant, S. (2000) *J. Mol. Biol.* **296,** 1319–1331.
11. Kolinski, A., Betancourt, M. R., Kihara, D., Rotkiewicz, P. & Skolnick, J. (2001) *Proteins* **44,** 133–149.
12. Kuriyan, J. & Cowburn, D. (1997) *Annu. Rev. Biophys. Biomol. Struct.* **26,** 259–288.
13. Meng, W., Sawasdikosol, S., Burakoff, S. J. & Eck, M. J. (1999) *Nature (London)* **398,** 84–90.
14. Schindler, C. & Darnell, J. (1995) *Annu. Rev. Biochem.* **64,** 621–651.
15. Harpur, A. G., Andres, A. C., Ziemiecki, A., Aston, R. R. & Wilks, A. F. (1992) *Oncogene* **7,** 1347–1353.
16. Yeh, T. C., Dondi, E., Uze, G. & Pellegrini, S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 8991–8996. (First Published July 25, 2000; 10.1073/pnas.160130297)
17. Bork, P. & Gibson, T. J. (1996) *Methods Enzymol.* **266,** 162–184.
18. Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996) *Methods Enzymol.* **266,** 383–402.
19. Yan, H., Krishnan, K., Lim, J. T. E., Contillo, L. G. & Krolewski, J. J. (1996) *Mol. Cell. Biol.* **16,** 2074–2082.
20. Gauzzi, M. C., Barbieri, G., Richter, M. F., Uze, G., Ling, L., Fellous, M. & Pellegrini, S. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11839–11844.
21. Ali, M. S., Sayeski, P. P. & Bernstein, K. E. (2000) *J. Biol. Chem.* **275,** 15586–15593.
22. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28,** 263–266.
23. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864.
24. Ponting, C. P., Schultz, J., Milpetz, F. & Bork, P. (1999) *Nucleic Acids Res.* **27,** 229–232.
25. Kampa, D. & Burnside, J. (2000) *Biochem. Biophys. Res. Commun.* **278,** 175–182.
26. Bradshaw, J. M. & Waksman, G. (1999) *Biochemistry* **38,** 5147–5154.
27. Mayer, B. J., Jackson, P. K., van Etten, R. A. & Baltimore, D. (1992) *Mol. Cell. Biol.* **12,** 609–618.
28. Bernstein, F. C., Koetzle, T. F., Williams, G. L. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
29. Yang, A. S. & Honig, B. (1999) *Proteins* **37,** Suppl. 3, 66–72.
30. Bairoch, A. & Apweiler, R. (1999) *Nucleic Acids Res.* **27,** 49–54.
31. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
32. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
33. Abagyan, R. A. & Batalov, S. (1997) *J. Mol. Biol.* **273,** 355–368.
34. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L., Jr. (2000) *Proteins* **40,** 6–22.
35. Sali, A. & Blundell, T. L. (1993) *J. Mol. Biol.* **234,** 779–815.
36. Luthy, R., Bowie, J. U. & Eisenberg, D. (1992) *Nature (London)* **356,** 83–85.
37. Eisenberg, D., Luthy, R. & Bowie, J. U. (1997) *Methods Enzymol.* **277,** 396–404.
38. Kuriyan, J. & Darnell, J. E., Jr. (1999) *Nature (London)* **398,** 22–23.
39. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998) *Bioinformatics* **14,** 892–893.
40. Rost, B., Sander, C. & Schneider, R. (1994) *Comput. Appl. Biosci.* **10,** 53–60.
41. King, R. D., Saqi, M., Sayle, R. & Sternberg, M. J. (1997) *Comput. Appl. Biosci.* **13,** 473–474.
42. Mattsson, P. T., Lappalainen, I., Backesjo, C. M., Brockmann, E. & Lauren, S. (2000) *J. Immunol.* **164,** 4170–4177.
43. Bashford, D. & Karplus, M. (1990) *Biochemistry* **29,** 10219–10225.
44. Yang, A. S., Gunner, M. R., Sampogna, R., Sharp, K. & Honig, B. (1993) *Proteins* **15,** 252–265.
45. Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1996) *Biochemistry* **35,** 7819–7833.
46. Sham, Y. Y., Chu, Z. T. & Warshel, A. (1997) *J. Phys. Chem.* **101,** 4458–4472.
47. Alexov, E. & Gunner, M. (1997) *Biophys. J.* **74,** 2075–2093.
48. Waksman, G., Kominos, D., Robertson, S. C., Pant, N., Baltimore, D., Birge, R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., *et al.* (1992) *Nature (London)* **358,** 646–653.
49. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.